DOCUMENT RESUME

ED 053 582 FL 002 273

AUTHOR	Ife, Anne E.; Standish, Peter
TITLE	Advanced-Level Testing of Foreign Language Proficiency: An Interim Report of the Post A-Level Spanish Project.
INSTITUTION	Essex Univ., Colchester (England). Language Centre.
PUB DATE	Jan 71
NOTE	36p.; Occasional Paper 11
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Achievement Tests, Charts, *College Students, Item Analysis, Language Ability, *Language Skills, Language Tests, Modern Languages, Objective Tests, *Spanish, *Student Evaluation, Tables (Data), *Test Construction, Tests
IDENTIFIERS	England, Palspan Project

ABSTRACT

This is a preliminary report concerning the development of tests which measure the linguistic ability in Spanish of English students at the beginning of their post A-level courses. The Palspan (Post A-level Spanish project) pilot test battery is comprised of five sub-tests of between 45 and 90 minutes in length which test for speaking, listening, reading, and writing skills. Processes involving marking and validation of the tests are discussed with particular reference to: (1) item analysis, (2) histograms, (3) variances and standard deviations, (4) correlation between tests, (5) relationship between test results and A-level results, (6) relationship between Palspan results and other external data, and (7) student performance. This study is heavily documented with linguistic and statistical data. (RL)

ED053582

University of Essex

Language Centre

OCCASIONAL PAPERS

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

FL002273

Colchester, January 1971

11

POST
A-LEVEL
SPANISH
PROJECT

Advanced-level testing of foreign language proficiency: an interim report of the Post A-level Spanish Project.[1]

Anne E. Ife
Peter Standish

The Post A-level Spanish Project (Palspan) was set up to devise language teaching materials at an 'advanced' level; that is, for use in those universities and colleges already closely involved in the teaching of language skills, and also for use in those universities or colleges which would like to stress these skills more were it not for the time and thought needed to provide suitable materials for the purpose. In the absence of a language-teaching specialist, language teaching is usually left to someone whose prior commitments in other fields force him to resort to translation and essay as the most familiar methods available.

To ensure its own relevance and usefulness the project needed to know in some detail what could be expected of the students in question, as regards linguistic ability in Spanish, at the beginning of their post A-level courses. The only information available was their A-level result - a single overall grade in an examination which in any case relies on translation and essay as means of language testing and which would appear to take into account none of the recent developments in language teaching and linguistics. Re-testing was therefore necessary, yet no existing tests were suitable.[2] It was thus decided to experiment with new tests, in the hope not only of gleaning useful information but also of producing ultimately a series of valid language tests for use at this level. This is preliminary report on those experiments.

The problem of language testing was approached with three major assumptions: first, that knowledge of a language involves a number of partially independent skills which can to a certain extent be tested separately; secondly, that those skills can be broken down into a number of specific structural points which can be tested in relative isolation (although this in itself can lead to the exclusion of creativity on the part of the student and needs to be compensated for in some way); and finally, that language differs according to context or situation, and according to region, and students should be aware of this.

The Palspan pilot test battery comprised five sub-tests of between 45 minutes and 1½ hours in length, and characterised more or less discretely by the following skills:

Test A - Speaking
Test B - Listening
Test C - Reading/Writing
Test D - Reading
Test E - Writing

A summary of these tests is to be found in Appendix 1. By way of explanation of the considerable total length of the tests, it was thought best to experiment with too much material initially, assuming that some of it would prove invalid, especially since there is no real precedent for testing at this level,[3] and as some of the techniques used were highly experimental.

An attempt was made to counterbalance objective techniques, which tended to isolate individual linguistic structures, by questions where those structures worked together in synthesis. Test material was drawn from varying sources and topics, and both American and Peninsular voices and sources were used. After completing tests A to E, candidates were invited to add personal details (though they remained anonymous), and to give opinions concerning, in particular, the administration and the difficulty of the tests. This information is summarised below.

In undertakings of the kind described here it is clear that the larger the sample, the more valid the results. The reliability of any kind of statement regarding normal student performance, whether overall or on particular sub-tests, is dependent on the size of the sample providing the data upon which such conclusions are based. In the present case, time and, more important, the limited sector of the population which is the concern of this study, set their own limits on the size of the sample. All universities and colleges in Britain known to accept students into degree courses with an A-level or equivalent knowledge of Spanish were contacted and invited to participate. The co-operation was encouraging, although administrative problems brought the withdrawal of some offers of collaboration.[4] The aims of the testing project were outlined to students who were asked, but not forced, to take the tests. Numbers taking particular tests varied between 50% and 60% of an estimated total population of 500.

In considering the control of testing conditions and the need to minimise the effect their variation might have on results, two alternative procedures were debated. If the collaborating centres, being fully briefed, conducted the tests themselves, it might be possible to ensure that all candidates take the tests at or near the same time, namely on entry to their courses. The advantages of such an approach are self-evident; alternatively, members of the project team could administer the tests, reducing technical and administrative

difficulties, and guaranteeing near parity of conditions during the tests themselves, though their scheduling in relation to each other and the dates on which they were taken would inevitably vary. The latter alternative was chosen; it proved expensive in time and money, but imposed a minimum burden on staff in collaborating centres, and afforded first-hand experience of student reactions.

While some students wrote appreciatively of the thoroughness of the tests, it is clear from many comments, and from the scores, that the test battery was too long - a fact foreseen when the pilot version was composed, as already mentioned. A drop in efficiency or enthusiasm seems to have affected some candidates. In the light of the test battery's first run, some material is shown to be impractical for technical reasons in all but ideal conditions; validation of individual items (outlined below) reveals some further material as suspect, and more as valid but redundant since it duplicates evidence produced reliably elsewhere. More detail about this is given below. The revised version of the test battery will have a total length of approximately half that of the pilot version.

It was again predictable that complaints would be registered over the use of non-Castilian voices, and subject matter in which students normally have no foreign language preparation. As has been emphasized before, the coverage of several varieties of Spanish was fundamental to the construction of the tests. It was recognised that most British students learned Castilian Spanish and that, at best, their contact with the language personally and culturally was very limited. But an important factor in discriminating between students is knowing something of their ability to handle different linguistic contexts, their adaptability to varying forms of language contact; if language learning involves four skills it may also be said to involve a variety of styles and subject matter.[5]

For the rest, there seems to have been no significant consensus of student opinion; comments tended to relate to time allowed for particular tests, to occasional lack of clarity in the rubrics, and to other administrative details. However, it may be worth recording the indignation of a few at being expected to have any oral/aural ability after A-level.

Marking and Validation Processes.

At the outset, selecting test material, devising techniques and drafting marking schemes were inevitably matters of personal judgment. It is possible, however, to evaluate some of these decisions in retrospect by the processes outlined below. It has already been mentioned that the test battery contained both objective and subjective tests. Objective tests could easily be marked right or wrong, though in multiple-choice

items the marking scheme included a correction for guessing. In some instances this was essential, since the choice involved only two alternatives, but many of the multiple-choice items had more than one distractor so that the guessing factor was in any case considerably reduced.

Some attempt was made to reduce the subjectivity of the other components, principally the production tests, both oral and written, by abstracting distinguishable aspects of production such as grammar, vocabulary, content, pronunciation, fluency and style, and marking each separately on a 5-point scale, ranging from 'poor' to 'excellent'. On a reading exercise certain pre-determined aspects of pronunciation were listened for, and pronunciation, intonation, and phrasing were each marked on a 5-point scale. In the re-telling of a story, certain fixed points of content were sought. Marking of the subjective tests proved a lengthy and somewhat uneconomical procedure, partly because of the attempt to increase the reliability in their marking.

The procedures to determine the internal validity of the test battery included the following:-[6]

1. Item Analysis.

For all the objective tests an item analysis was undertaken to assess the efficiency of each individual item. A successful item is one which is neither too difficult nor too easy, and which draws from each candidate a performance corresponding to his performance on the test as a whole. For each of the tests B, C, and D, scripts were sorted into rank order according to individual overall scores on that particular test, and then divided into three groups; an upper, a middle, and a lower group, with always the same number of scripts in the upper and lower groups. For each item a record was kept of the number of students in each group giving the correct answer, and on the basis of this information facility values and discrimination indices were calculated.[7]

The facility value of an item shows its difficulty in terms of the number of candidates giving the correct answer. In general, a facility value of between .3 and .7 (i.e. an item with between 30% and 70% correct answers) would indicate a successful item, though a certain amount of leeway at either extreme might be allowed. The discrimination index of an item shows how well it reflects the overall distribution of scores on the test as a whole; that is, how many more people in the upper group than in the lower group give the correct answer. An item can be said to discriminate well if it has a discrimination index of .3 or more. An item with a negative or zero discrimination index is obviously not functioning properly.

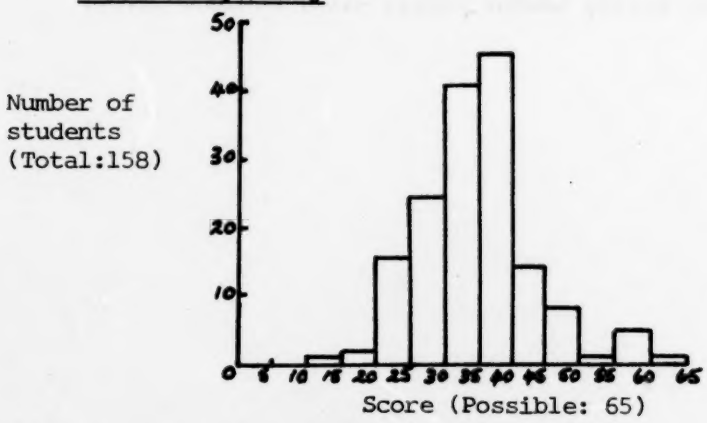
For those items with distractors, an analysis of these was made to find out whether they distracted too much or too little. This was done simply by recording the number of students choosing each distractor. If done with scripts divided into three groups, as above, additional information can emerge, (for instance, all students in one group might opt for the same distractor, while those in another group choose a different one).

A detailed report of the various facility values and discrimination indices would mean little here without a full account of each item. Suffice it to say that a fairly large proportion of items need to be dropped or changed in some way: an estimate of 50% would not be too high. One can only repeat again that the test battery was originally made very long and unwieldy precisely because it was realised that much would need changing at this stage. In many cases the fault in an item lies with a badly chosen distractor and can easily be rectified. In other cases the fault is caused by the fact that the question is testing a feature of the language which students have mastered without difficulty. This applies, for instance, to a section concerned with the rules of stress.

2. Histograms.

The series of histograms below shows the shape and range of the distribution of overall scores for each of the 5 tests.[8] Scores given on the histogram itself are raw scores. The columns of the histogram represent the number of students (shown on the vertical axis) having a score within the range indicated along the horizontal axis. Accompanying each histogram is a set of scores, including the highest, the lowest and the median scores actually achieved, together with the mean score.[9] To make them more meaningful these scores are given as percentages.

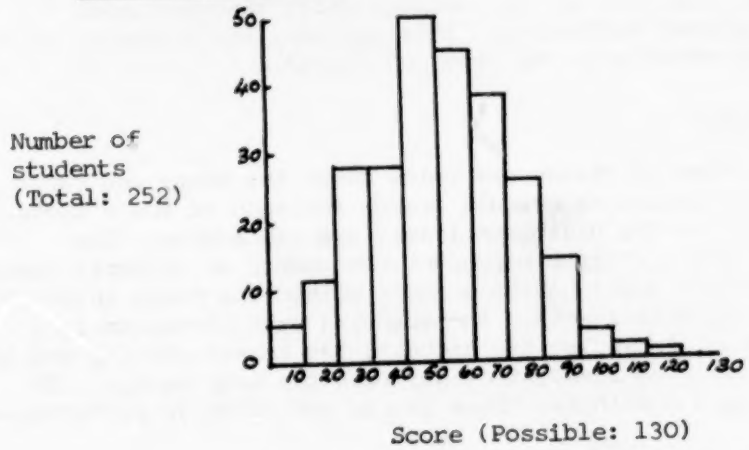
Test A - Speaking



Highest score: 100%; Lowest score: 21.6%; Median: 53.9%; Mean: 54.2%.

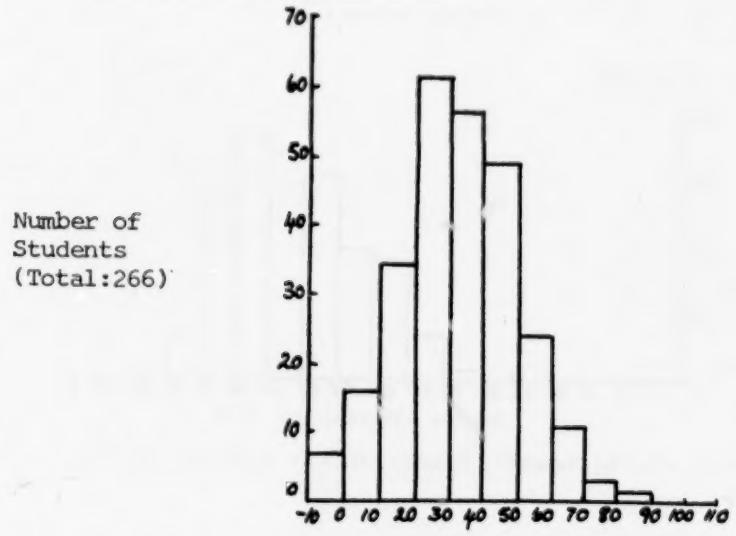
Because of administration difficulties, and a number of spoiled tapes, the final number of students supplying a completed tape was only 158. The one score of 100% was that of a native speaker. Otherwise the marks are spread over a wide range, though with a considerable bunching around the mean; just over half the total number of students scored between 46% and 61%, yet only 18% of them scored more than that. This was one of the subjectively marked components of the test battery and the shape of the distribution curve may well reflect the difficulty in discriminating between students of unexceptional ability in a test of this kind, in spite of the efforts made to reduce the subjectivity. A more liberal use of the upper and lower limits of the 5-point marking scale may be needed, thus allowing more freedom for discrimination in the middle ranges, whereas the temptation is to reserve these marks for outstandingly good or bad candidates.

Test B - Listening



Highest score: 85.5%; Lowest score: 4.6%; Median: 39.2%; Mean: 39.3%.

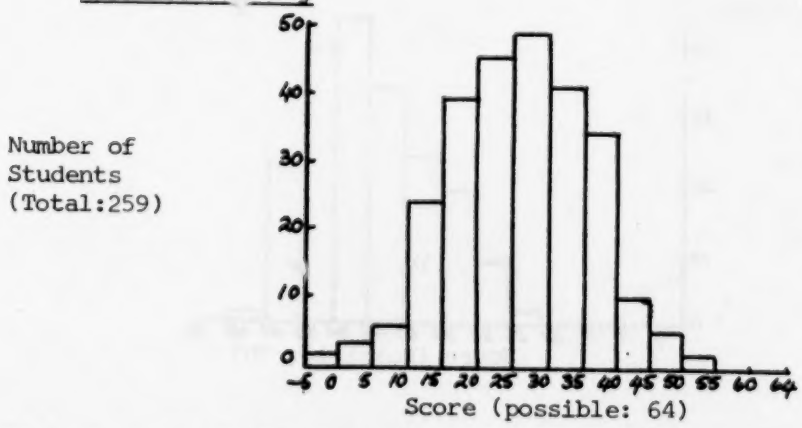
Test C - Reading and Writing



Highest score: 81.8%; Lowest score:-0.82%; Median: 30%; Mean: 30.8%

The negative lowest score is accounted for by the deduction for guessing in multiple-choice items. (The same applies in test D). Undoubtedly this test proved very difficult: scores are concentrated in the lower half of the score range and the mean is only 30.8%. Nevertheless, the range of scores is very wide.

Test D - Reading

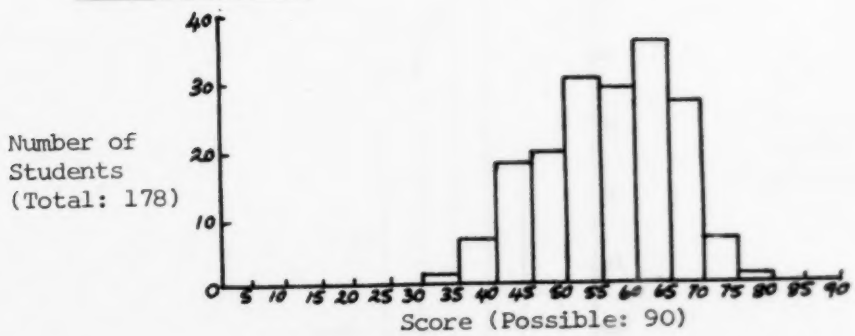


Highest score: 82.9%; Lowest score:-7.8%; Median: 40.6%; Mean: 41.7%.

The item analysis showed Test D to be easier than the

histogram would suggest. However, the deduction for guessing compensates for this to a certain extent.

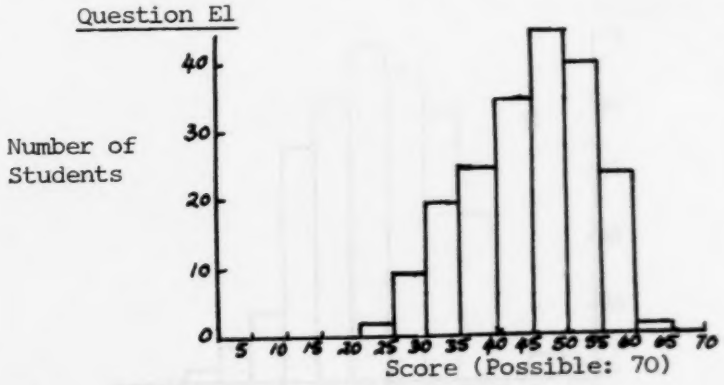
Test E - Writing



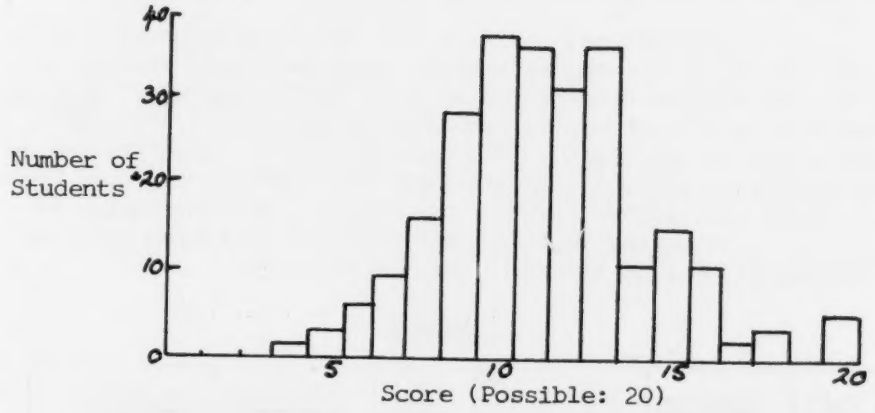
Highest score: 85.5%; Lowest score: 36.6%; Median: 63.4%; Mean: 63.3%.

As with test A the total number of students is unfortunately low, in this case because an unclear instruction rendered several papers invalid. The distribution curve reveals problems with this test; firstly, the range of scores achieved is very narrow, with everyone scoring within a range of 50% (compared with 80% to 90% for all the other tests), which means that it is not a very successful discriminator. Furthermore, it is also too easy, as scores are concentrated in the upper half of the range and the mean is very high. A separate histogram for each of the questions which comprised E soon reveals the main cause of this.

Question E1



Question E2



Clearly, the fault lies with Question E1 which provided little difficulty for students of this standard: the bulk of scores for this question are at the upper end of the mark scale. On the other hand, the range for E2 is good, although the distribution of the marks again reveals the problem found with subjective marking even with a clearly delimited scale, as in Test A.

Of the five tests the best discriminators would seem to be the objective tests B, C and D, where the distribution curves show both wide score ranges, and a steady increase in distribution from the lowest to the median scores and a steady decrease from there to the highest score. Test A achieved an acceptable range of scores but the distribution within that range could be better. There is a particularly sharp drop in numbers above the 36-40 score range. Test E on the other hand shows a fairly even distribution but reveals the other problems already noted.

Normally a histogram would be produced for the test battery as a whole, but in this case, not only were the sub-tests quite discrete entities, but there were also considerable difficulties in producing a set of overall results, when numbers were so depleted for two of the tests and when only 84 students completed all five tests. Overall percentages were produced for the benefit of the candidates with the proviso that these were not very reliable, mainly because a percentage based on the easier tests alone would appear better than one based on the more difficult tests. For interest, however, the range of scores on the whole test battery was as follows:-

Highest score: 78.1%; Lowest score: 10.8%; Mean: 45.5%.

Despite the caution appropriate here,these scores suggest that the tests as a whole discriminate reasonably well between students.

3. Variances and standard deviations[10]

For each test the variance and the standard deviation were calculated, as a means of showing more precisely some of the information presented graphically in the histograms - that is, how well each test distributes the candidates.[11] This will be explained further below. The results of the calculations are presented in Table I together with the maximum score and mean for each test (given here as raw scores), and the number of students, since the variance and standard deviation will be meaningful only if seen in relation to these.

TABLE I

Test	Maximum score	Number of students	Mean	Variance	Standard deviation
A	65	155	35.22	71.54	8.46
B	130	236	51.24	422.56	20.56
C	110	261	33.83	286.06	16.91
D	64	253	26.70	97.55	9.88
E	90	170	56.89	91.09	9.54

If the standard deviation is added to the mean and then subtracted from the mean, the two resulting scores are the extremes of a distance along the score scale within which the scores of two thirds of the candidates tested can be said to fall. The greater that distance the better the test spreads the population. This means, for instance, that in Test A two-thirds of the candidates scored between 26.8 and 43.7 out of a possible total of 65; that is, between 41% and 67%. (The figure of 26.8 is achieved by subtracting the standard deviation for Test A from the mean for A, and 43.7 by adding the two). By the same process, the ranges for the same proportion of the population for each of the tests B, C and D are rather wider; between 24% and 55%, 15% and 46%, and 26% and 57% respectively. That for E is narrower: between 52% and 74%.

4. Correlation between tests

All the tests were considered to examine different skills, with the exception of C which tested a combination of those skills tested in D and E. One would therefore expect low correlations between all tests except C and D, and C and E. In practice, problems occurred in finding correlations, mainly

because only 84 students completed all five tests, as mentioned. Correlations could have been calculated on the basis of all the results available, but with the computer programme used there was no way of substituting satisfactorily for missing values in those cases where students had taken some, but not all, of the tests. (It is hoped that this will be rectified by next year.) Simply to have substituted the mean score for those missing values would have resulted in a very suspect set of correlations; reducing the sample to those 84 students with a full set of scores also meant unreliable results, though less so than with the other method, and on the whole this approach seemed preferable. The results obtained are given in Table 2.

TABLE 2

Test	A	B	C	D
B	0.503			
C	0.528	0.496		
D	0.442	0.470	0.647	
E	0.338	0.283	0.527	0.342

E is shown to have a low correlation with the independent tests A, B, and D, but a higher one with C, as one would expect. Surprisingly, A shows a high correlation with all the remaining tests, as does B. However, the high correlation between C and D is as predicted. It must be emphasized, though, that these correlations could well have been different had the sample been larger and therefore too much significance should not be placed on these results.

5. Relationship between test results and A-level results

In order to ascertain how far results on the Palspan tests could be predicted by A-level results, students were asked to supply details of their A-level examinations. Some 80% of candidates tested had an A-level qualification (including Scottish and Irish equivalents), the remainder being accounted for by those who had completed additional one-year courses preparatory to degree work, and by a handful of non-British students (with some native speakers among them). Of the A-level students, approximately 26% had achieved grade A, 43% grade B, 16% grade C, 8% grade D and 7% E or lower.[12] Information about A-level was classified in seven categories, the first five to cover the five grades A to E (this latter including three people with O or F grades), the sixth to cover those for whom no grade was known, and the seventh those who had begun Spanish only at

TABLE 3

	A-level Grades					Not known	University only
	A	B	C	D	E		
Test A	+1.85	+1.11	-2.61	-2.32	-3.37	+4.68	-0.60
No. of students	(30)	(51)	(23)	(10)	(18)	(10)	(13)
Test B	+6.20	-1.34	-7.09	+3.26	-9.84	+2.39	+6.35
No. of students	(48)	(78)	(33)	(16)	(20)	(19)	(22)
Test C	+9.50	+0.35	-6.14	-1.69	-17.73	+2.86	-1.04
No. of students	(55)	(94)	(36)	(14)	(20)	(23)	(19)
Test D	+3.20	+0.74	-2.24	-1.70	-8.00	+0.36	+0.62
No. of students	(55)	(95)	(35)	(14)	(20)	(18)	(16)
Test E	+3.28	+1.16	-1.81	-2.26	-7.36	+0.77	-1.96
No. of students	(34)	(59)	(23)	(11)	(13)	(15)	(15)

university level. The results are shown in Table 3. Figures in brackets represent the number of students in each category, as an indication of the size of sample on which the results are based. The other figures represent the amount by which the mean score of students within the various categories differs from the overall mean for the relevant test. To appreciate the full significance of these results it is important to bear in mind Table 1 giving maximum and mean scores and standard deviations. The results which differ significantly from each other are circled.[13]

The chart shows that for the oral test, A, there was no statistically significant difference between the results achieved by students with different grades at A-level. This was perhaps to be expected. For the other four tests, however, the differences were statistically significant, and in test C the differences were particularly striking; students with grade A at A-level scored on average 9.5 marks more than the general mean for test C, whereas those with grade E scored on average 17.73 less than the mean. (There was even a considerable difference in performance between students with grades A and B). It may be noticed that on all tests except E, students with grade D did better than those with grade C; however, in no case was this difference statistically significant.

Table 4 gives a summary of the degree to which A-level can be said to predict the results achieved on the Palspan tests.[14] The degrees are given in terms of percentages of the variance (or total sum of squares) of the tests, which are directly attributable to differences in A-level performance. 10% could be considered to indicate a close relationship.

TABLE 4

Test	% of variance attributable to A-level
A	7.8%
B	6.8%
C	17.2%
D	8.6%
E	8.7%

This summary serves to emphasise the very close relationship indicated between test C and A-level, a surprising result, not least because test C was in many ways the most experimental and the least like A-level examinations in format. On the other hand, C did test a combination of reading and writing skills and in this respect resembled the A-level language tests closely. The relationships indicated between A-level and the other four tests are all very similar; one might perhaps have expected a wider gulf between the oral/aural tests and A-level than is indicated here. One interesting point is the fact that test C proved the most difficult and it is precisely this test which seems to bear the closest relationship to the G.C.E. exam.

6. Relationship between Palspan results and other external data

Students were also invited to supply information about the time they had spent in a Spanish-speaking country, and about the length of time during which they had studied Spanish. No less than 57% of all candidates had spent two weeks or more in a Spanish-speaking country for purposes which they described as not solely pleasure; the figure rises to 68% for candidates with grade A at A-level and drops correspondingly for lower grades. However, of the total number of candidates having been abroad by far the greater number had stayed for relatively short periods of time: 54.7% had stayed for 2 to 4 weeks; 28% for 5 weeks to 3 months; 6.7% for 4 to 6 months; 8% for 'frequent visits' and 2.6% for 10 months to 2 years.

With regard to length of study, a somewhat surprising proportion of 76% of A-level students had studied Spanish for more than 2 years, but here there was no notable increase (1%) in the proportion when calculated for A and B grades only.

Information of this kind was classified in six categories which included the following: bilingual students (i.e. British students brought up in Spanish-speaking countries); those who had studied Spanish abroad for any length of time and, in another category, those who had not done so; those who had studied Spanish for 4 years or more; those with 3 or less years' study; and finally, a category for all students for whom none of these details were known. Students with university experience only were accounted for elsewhere. Table 5 is based on a statistical analysis of the computer output, and shows how far each of these variables affected the test results. It is produced on the same basis as Table 3 showing the relationship of A-level grades to the test results and again, to be fully appreciated, the chart of maximum scores, mean scores and standard deviations (Table 1) needs to be borne in mind: numbers in brackets indicate the number of students in each category; figures preceded by a + or - sign indicate the amount by which the mean score for students in the various categories differed from the overall mean score for a particular test. Statistically significant differences

TABLE 5

Test	Bilinguals	Abroad	Not Abroad	4 years +	3 years -	Not known
A	+24.53	+1.92	-4.76	-0.35	-0.94	+12.22
No. of Students	(4)	(86)	(51)	(86)	(51)	(1)
B	+27.76	+3.99	-10.77	-1.78	-0.35	-7.52
No. of Students	(6)	(129)	(72)	(128)	(73)	(7)
C	+30.42	+2.15	-6.61	+2.13	-7.15	+0.47
No. of Students	(8)	(143)	(81)	(148)	(76)	(10)
D	+6.80	+1.06	-2.49	+0.73	-2.24	+1.86
No. of Students	(4)	(142)	(82)	(151)	(73)	(9)
E	+12.11	+0.71	-2.0	-0.12	-0.44	+1.48
No. of Students	(4)	(94)	(49)	(99)	(44)	(8)

between categories are again circled. Where one figure only is ringed, this indicates a significant difference between the performance of students in the relevant category and that of students in all the other categories, for that particular test only; where more than one figure is ringed, then a significant difference in performance between all categories within that ring is indicated.

The main function of the few bilingual students was to serve as a control, and the results show them to hold the advantage one would expect on all except the reading-comprehension test; test D was the only case where a statistically-significant improvement in performance was not obtained from them. The most likely cause of this is that the test was not sufficiently demanding (a fact borne out by the validation of individual items) for any inherent advantage, such as bilingualism, to make any difference.

Students with experience of study abroad performed considerably better than those without such experience on all tests except E, the writing test. This would be predictable were it not for the fact that in most cases the length of the stay abroad was very short - only 4 weeks or less in more than half the cases. That such a short stay appears to be so beneficial is both surprising and pleasing.

Length of study was reflected only in the results of tests C and D, where there was an appreciable difference in performance between those with 4 or more years' study and those with less than 3 years.

The effect of these three variables (bilingualism, study abroad and length of study) is more clearly seen in the summarised chart - see Table 6. As with the earlier summary of the influence of G.C.E. results, the influence is shown in terms of a percentage of the variance, and again, 10% can be taken to indicate a close relationship. Statistically significant results are circled.

TABLE 6

Test	Percentage of variance, attributable to:			
	Bilingualism	Study abroad	Length of study	Other
A	(22.4%)	(13.0%)	0.1%	1.2%
B	(4.8%)	(10.0%)	0.1%	1.5%
C	(10.2%)	(5.3%)	(5.8%)	0.05%
D	0.8%	(2.7%)	(1.8%)	0.2%
E	(3.9%)	1.5%	0.05%	0.4%

That bilingualism should have the greatest effect on oral work was predictable, as was the fact that study abroad should influence most the oral/aural skills. Although not predictable, it is not so surprising that length of study should have no great effect; so little emphasis is given to oral/aural skills in pre-A-level teaching that length of study could hardly make a great deal of difference. On the other hand, the two tests which did appear to be affected by previous study were concerned with precisely those skills (that is, reading-writing, and reading) which receive most attention in schools and here one would expect to see some difference in level of achievement between students with greater or less experience. However,this also applies to test E, the writing test: writing receives even more attention than reading in most schools. The absence of any effect of length of previous study on test E may, therefore, be an indication of the tremendous effort given by teachers of short A-level courses to bring pupils up to the standard of others with much longer experience of writing Spanish.

This concluded the validation processes for the initial testing programme, from which many areas of interest have emerged. The test battery has now been revised and modified on the basis of information gained from the validation, hopefully eliminating all points of weakness, and it will be readministered in the academic year 1970-1. It will be interesting to compare the relationships shown here between the Palspan tests and A-levels and other external data with those that will emerge when the test battery has itself been shown to be more internally reliable. Next year's testing programme should be another step in that direction. Meanwhile, although the results presented here can only be provisional and inconclusive, it is hoped that they will, nevertheless, be of interest.

Student performance.

If the purpose of the initial testing programme was primarily the validation of the test itself, it was also hoped that particular areas of student error would become apparent. Undoubtedly, certain errors were caused by bad test construction rather than by lack of knowledge on the part of the testees, and care has been taken not to attach too much significance to isolated incidence of mass error. The following remarks are based on observation of the types of error that recur often enough to suggest a need for some remedial teaching.

Oral work left much to be desired. Pronunciation was usually poor; students were either unaware of, or careless in using, Spanish sound patterns, in particular the quality and distribution of allophones (and one suspects that this is the result of ignorance rather than carelessness); few were aware that Spanish, unlike English, does not observe word boundaries within breath groups, and therefore produced a careful hiatus

between every word; intonation was poor in exercises involving oral composition, but showed a marked improvement on the reading passage; on the other hand few found difficulty with stress patterns within words. The language produced was generally punctuated by long silences, false starts and unfinished sentences which were excessive even if one allows that a perfectly fluent and well-formed piece of oral composition would be unlikely even in the native language. When full sentences were uttered they were for the most part simple, while the lack or misuse of varied or relatively complex linking expressions gave an impression of lack of cohesion and sophistication to the whole, as did the limited vocabulary choice.

To expand further on these points: with the exception of the native and bilingual speakers and a few more able students, pronunciation was sometimes good in parts but inconsistent, but more often consistently poor. The weaknesses came in expected areas: vowels were too long and sometimes too centralised, tending towards the English /ə/ in unstressed syllables, although the diphthongisation that is a temptation to English-speakers was usually carefully avoided; many students seemed completely unaware of the fricatives [β], [ð], and [ɣ] in intervocalic and other positions and consistently produced the related stops, while others produced the dental fricative sporadically, mainly in common words like nada, yet not in past participles, and did not produce the bilabial or velar fricatives at all; a good many did not know that orthographic v is pronounced [b] or [β] according to environment and regularly produced a voiced labio-dental fricative which Spanish does not possess; the stops [b] and [p] and [t̪] and [d̪] were as a rule given too much aspiration; it was pleasing to note that some people made very successful attempts at the apico-alveolar [s], but again, a minority; also, some good alveolar trills, [r], were produced but unfortunately usually indiscriminately, without allophonic variations; not many produced the phonemically distinct flap [ɾ]; finally,there were rarely any variations of the nasal /n/ in the light of a following bilabial, dental, labio-dental or velar.

Sophistication in expression is obviously something which has to be acquired in post A-level teaching; few embedded clauses were used, with the exception of relative clauses, though this may be to a certain extent a feature of informal speech in general. The conjunctions y and pero were vastly overworked, and no attempt was made to use such useful linking expressions as no obstante, mientras tanto, sin embargo, a pesar de etc. There was a common and strange use of pues (presumably by analogy with the French puis) where después, luego or entonces would have been more appropriate.

Several problems arose concerning verbal cohesion. A considerable uncertainty about the use of the subjunctive

revealed itself in the avoidance of constructions involving subjunctives, in the abandoning of a sentence when the speaker realised that he was involving himself in subjunctive constructions, or simply in the failure to produce a subjunctive form of the verb when one was needed. Also, problems with tense were frequent. Often students did not maintain a unity of tense, especially in the cartoon commentary (question A2) where they hovered doubtfully between a present tense description of events and a past tense story, e.g.

'Hay un hombre que anda..... Entonces vio a una mujer.....'

More seriously, there seemed to be doubt as to the permitted sequences of tense, e.g.

'Sentí que voy a ser muy feliz aquí'

In the first person narrative, A3, several people made use of the immediate past tense where the preterite or imperfect was needed, again influenced no doubt by the French usage, e.g.

'He llegado aquí hace tres semanas.....'

And although the students seemed better able to talk about themselves than to describe the cartoon, they nevertheless were inept with first person verb forms, especially preterites. The whole question of tense and mood demands revision.

Ser and estar were inevitable thorns. Practice is needed too in concord of gender and number of nouns and adjectives, and of number of nouns and verbs. These problems occurred more in oral production than in written work, probably because one has more time for thought when writing, and obviously the rules for agreement have not yet become automatic anough for students to follow them rapidly and unconsciously.

Knowledge of prepositions was very weak, and constituted a serious hindrance to fluent speech, even more perhaps than ignorance of lexical items. Frequently a long hesitation was followed by the Spanish equivalent of the English preposition that would be used in that context, and too often one heard such items as:-

'Estudiaré aquí por cuatro años...'

'Cuando estaba a la escuela...'

This also extends to knowledge of which prepositions are associated with particular verbs. Surprisingly, few forgot the 'personal a'.

Vocabulary choice was extremely limited, in particular in

the cartoon commentary where it was predetermined to a degree by the pictures. There were too many literal translations from English, such as:-

'Tiene su marido con ella...'(Few thought of acompañar)

'Empuja un perambulator...'(Some did supply coche de niños or cochecillo)

'Tiene un perro en una cadena'

Other typical anglicisms were duro for difícil, atmósfera for ambiente, para empezar for al principio etc. The other problem was the interference of French and Latin in such cases as a causa de, the eternal pues and even el can for el perro.

It is more difficult to make generalizations about performance in aural work because here much depends on the validity of the tests. The transcription exercise, A1, had to be abandoned owing to difficulty in its administration. However, from observation of a few attempts at this question, and from noting student comments, it would appear to be too difficult an exercise. Those who actually attempted the transcription seemed totally unaccustomed to Spanish read at a normal speed and complained that the passage was read too fast to understand (although they had control over their tapes and could listen to the passage as often as they wished). There were also complaints that the language laboratory deprived people of a view of the speaker's mouth. Certainly, few managed to transcribe more than half the passage in the 20 minutes allowed, and even so, many gaps were left and oddities produced. The low level performance here casts serious doubt on the value of the traditional dictation exercise in Spanish which is seldom a difficult task.

Tests of aural discrimination of phonemic differences in words or groups of words (question B5) proved, on the whole, extremely easy. This may be the fault of the test items, but it could also be that the testing of phonemic differences is largely superfluous at this stage. Those contrasts which did seem to provide some real trouble were the following:-

/f/	: /θ/	.....	rifar	: rizar
/s/	: /θ/	.....	eso es	: es soez
/t/	: /d/	.....	este miedo	: es de miedo
/rj/	: /ʎ/	.....	tiene varias	: tiene vallas
/ɣ/	: /k/	.....	no rasgues	: no rasques

Allophonic differences presented more of a problem. Seemingly, the stop and fricative allophones of /b/, /d/ and /g/ gave most trouble in this test, particularly the items given below, in which students were required to state whether

the relevant allophones in each pair were the same or different:-

[b] : [β] ..... voz : dar voces
[β] : [β] ..... malva : llave
[ð] : [ð] ..... verde : nudo
[ɣ] : [ɣ] ..... el guiso : el gato

Aural discrimination of stress proved exceptionally easy, and, bearing in mind that students also had little difficulty in producing the correct stress patterns, testing of this feature will be disregarded in future. With the exception of one item, the intonation test was also too easy. This was probably a case of the test itself being badly constructed since all the items were very short, with clearcut intonation patterns and were recorded specifically for this test. Intonation contours are rarely so simple and straightforward and it is possible that if the items were not only longer but also taken from speech recorded naturally then they would provide a better gauge of the students' ability.

Aural comprehension was tested elsewhere in ways which were undoubtedly quite new to most of the candidates, particularly in questions B6, 7 and 8. The most demanding test of aural comprehension proved to be question B6, where the distractors played on the possible mis-perception of particular linguistic items or on the misunderstanding of certain idiomatic expressions. Question B7 was much less demanding, and it may well be that the guessing factor could account for the apparent simplicity of this test. Question B8 was unsatisfactory mainly from the point of view of finding a suitable marking scheme. It was marked on a basis of number of points accurately reproduced, maximum 25, thus making memory a decisive factor. Only 11 students scored more than 20 marks, but more disturbing was the fact that some 108 students of a total of 246 scored less than 10 marks, implying that they understood or recalled accurately less than 40% of the story they had heard, even after three readings. Some may not have realised how fully the story was to be re-told, but in other cases it was only too obvious from the garbled version that resulted that only one or two points of the original had been understood.

In general, and in spite of B7, which appears to tell a different story, students were far from confident where aural comprehension is concerned, in particular on tests which involved accurate interpretation of specific linguistic items (as in A1 and B6), rather than just an overall impression of the content of the recorded passage.

In the test of reading and writing, C, the first question was very successful, although difficult from the testees'

viewpoint. On only twelve out of twenty-five items did more than 50% of the candidates give a correct answer. Words to be eliminated were of three types:-

i) Words which, in context, were semantically impossible ('pandillas de zanahorias jóvenes').

ii) Words which were semantically possible but broke rules of concord or were morphologically incorrect in some way ('la buen serie..')

iii) Words which, loosely speaking, broke syntactic rules ('se los limitan a ridiculizar..', where los has no referent; or 'cuyo más frecuente en lugar de reunión..')

This last type, which interfered drastically with the meaning of the passage, in fact caused least difficulty, though of these the hardest were those items where two finite verbs or prepositions were adjacent to each other and a choice obviously had to be made. By far the most difficult were items of type (ii), that is, those interfering least with the meaning of the passage, and it is arguable that these should not be included here since C4 consists entirely of similar items.

The deliberate mistakes in C4 were almost all morphological (the one exception will be eliminated next year): they involved:-

i) Mistakes of concord.

ii) Mistakes of tense.

iii) Mistakes of mood.

iv) Mistakes of word formation (concepto for concebido).

v) Use of the wrong one of a pair of contextually determined variants (pero instead of sino).

vi) Omission of prepositions, articles, reflexive pronouns. (To be omitted next year)

Just as with the items of type (ii) in question C1, none of these interfered with the meaning of the passage to any great extent, and again students found great difficulty in identifying the mistakes. These, as we shall see later, are just the mistakes they themselves produced in written work, so that in this area they obviously find extreme difficulty in both recognising and producing accurate Spanish.

Question C2, in spite of much complaint from the students about the difficulty of the exercise, was nevertheless a very successful test. Items can be classified as follows:-

i) Part of a word was given, normally the stem, so that the meaning was self-evident, as was the class of word required. The testees were required to supply either the

prefix or the suffix.

ii) The word to be supplied demanded a grasp of the grammatical construction being used and was normally a function word.

iii) One part of a compound expression was missing and completion demanded a knowledge of that particular idiom (no puede menos, or así como.. etc.).

None of these types presented any more difficulty than the others. In each case, roughly half the items were difficult, the others easy, so it is difficult to draw from this any conclusion other than that some items were better constructed than others.

The multiple-choice grammar items of question C3 were very easy. Of the items giving any kind of difficulty, all involved the use of verbs in one form or another: namely in the use of ser and estar, the imperative form, the sequence of tenses, and strangely, the use of hacer in referring to the weather. This is further evidence of the uncertainty indicated elsewhere in verbal constructions.

Judging from performance on test D, the standard of reading comprehension appeared to be high. All questions except D5 were too easy; surprisingly, a journalistic passage, rather than any of the others which dealt with philosophical and sociological topics or made use of very colloquial language, gave most difficulty. Standards of comprehension were underestimated and an attempt has been made to increase the difficulty of this component in the new test battery.

Problems revealing themselves in the writing test included many of those evident in the speaking test. E1, the reporting of the letter, involved mainly changes of tense and person. The former was achieved quite well, although with some uncertainty as to where a preterite was needed and where an imperfect. There were also a few peculiar attempts at pluperfect construction resulting in a 'double' pluperfect, e.g. 'había habido tenido..'. Again there were problems with sequence of tense, but apart from this the only difficulties concerned demonstratives, pronouns and verb forms to agree with a change of person.

E2 was a successful question but revealed a considerable lack of ability in writing a précis; candidates either produced a word for word translation of the English, failing to finish within the allotted time, or else omitted significant points in the passage, concentrated on trifling details and therefore lost marks for coverage.

Mistakes again involved verbal constructions: tense, the use of the subjunctive, the uses of ser and estar (in spite of all the emphasis given to this point in schools), and apart from

these, the particular difficulty with prepositions. In this exercise students rarely used reflexive constructions but, influenced by English, made use of passive constructions instead, which are not necessarily wrong but relatively infrequent in Spanish. Those who did use reflexive constructions seemed not to know what the verb should agree with, especially if a plural was involved.

Knowledge of vocabulary was weak. The passage did require a knowledge of the Spanish equivalents of seige, citadel, colonel, prisoner, capture, etc. but these are not as rare or technical as was claimed, and frequently very common words were either unknown or forgotten, the word periodista for instance. A confusion with groups of semantically similar words was evident, especially when they involved different constructions. One thinks immediately of carecer, faltar, and hacer falta.

Style, as was perhaps to be expected, was rarely very good, and in this area bilingual students excelled. Some made creditable attempts to produce a piece of prose which read authentically, but most were too concerned with trying to convey the basic facts to give much thought to stylistic devices. Ideas thought out in English were obviously translated word for word into Spanish, often with strange results. Apart from the usual mistranslations and analogies with other foreign languages, some of the 'Englishness' of their style, at least, could be eliminated by a greater use of reflexive and active constructions and less of passive ones, by giving more thought to word order, and by learning some idiomatic linking expressions, as for instance por eso (instead of producing cumbersome equivalents to English or French, like 'esto lo hizo hacer....' or 'a causa de esto...'. This, at least, should be possible by A-level standard.

University of Essex
August, 1970.

NOTES

1. The project is a joint undertaking by the Language Centre, University of Essex, and the Department of Spanish, Queen Mary College London, and is supported by a grant from the Department of Education and Science. Co-directors of the project are Miss L.E.Ingamells, of Queen Mary College, and Mr. P.Standish, of Essex. Research Officer is Mrs. A.E.Ife, of Essex.

2. Most published tests were too elementary. The nearest, as regards level of achievement, is the Modern Language Association Foreign Language Proficiency Tests for Teachers and Advanced Students. However the design and content of this test were considered unsuitable for our purposes.

3. A similar, but more limited, testing programme in German has been undertaken by Mr. W. Grauberg and Dr. R. Hartmann, of the Language Centre, University of Nottingham.

4. Details of collaborating centres appear in Appendix 2.

5. The primary aim of the Palspan tests is not to provide an alternative means of assessing [illegible] A-level syllabus, but to discriminate between a given s[illegible]e of entrants to degree courses.

6. Thanks are due to Dr. A. Pilliner of the University of Edinburgh who, as testing adviser to the project, has given invaluable advice and information about methods of test validation, and also test construction.

7. $\text{Facility Value} = \frac{S1+S2+S3}{N}$ where S1, S2, S3 refer to the number of students supplying a correct answer in the upper, middle and lower groups respectively, and N is the total number of students taking the test.

   $\text{Discrimination Index} = \frac{S1-S3}{N/3}$ where S1, S3 are as above, and N/3 is the number of students in the upper and lower groups (necessarily the same).

8. On histograms in general and other statistical terms used see, for example,P. Blommers and E.F. Lindquist, Elementary Statistical Methods in Psychology and Education, University of London Press, 1965, where they are discussed with particular reference to educational testing.

9. Blommers and Lindquist (op.cit.) give the following definitions of median and mean:

'The median of a distribution is the point on the score scale below which one-half, or 50%, of the scores fall.'
'The mean of a distribution of scores is the point on the score scale corresponding to the sum of the scores divided by their number.'

10. The variance is achieved by adding the squares of the differences of individual scores from the mean, and then finding the mean of the resulting sum. From the variance the standard deviation is derived, this being the square root of the variance. The standard deviation can be interpreted as a distance along the score scale.

11. The computing of results and the statistical analysis of the computer output was very kindly carried out by Mr.I.Russell, of the Department of Mathematics, University of Essex.

12. This confirms that the effective minimum entry qualification is normally grade B for universities. There is also some evidence that certain institutions admit students with better A-level grades than others, though numbers in these cases are small and not wholly reliable; it is quite possible, for instance, that it was the best, most confident students who chose to take the tests in centres where there was no pressure at all to do so.

13. The basis for the statement that these results are statistically significant is the analysis of variance, full details of which can be found in Appendix 3. To quote a result as statistically significant implies 'that the observed difference differs from zero by an amount greater than can reasonably be explained in terms of random sampling fluctuation' (Blommers and Lindquist).

14. The chart is a summary of the analysis of variance given in Appendix 3.

APPENDIX 1

The following is a brief description of the Palspan test battery, as administered during the first term of the academic year 1969-1970.

Tests A and B ..... Listening and Speaking (1½ hours)

Tests C, D and E ..... Reading and Writing (3 hours)

Test A

Tested mainly oral skills, with the exception of one aural test which was included here to save involving students with more than one tape.

Included:-

1. A recorded and unedited news item to be transcribed by the student, who had control over his tape for 20 minutes and could listen to the text as often as he wished.
2. A cartoon sequence to form the basis of a 2-minute recorded commentary.
3. A fixed topic on which to talk for 5 minutes, after time given for preparation.
4. A passage of Spanish to read and record.

Test B

Tested aural skills. All questions except 8 were objective.

Included:-

5. Tests of discrimination of phonemes, allophones, stress and intonation.
6. Aural comprehension - choice from 3 alternatives of a suitable written answer for a recorded question.
7. Aural comprehension - involved stating, in terms of true (T) or false (F), the validity of written statements based on a recorded passage.
8. A short contemporary story (about 3 minutes long), at normal speed, with opportunity for note-taking between the second and third readings, to be re-told, on paper, in English.

Test C

Tested a combination of reading and writing skills.

Included:-

1. A passage of Spanish (serious newspaper article), with words inserted for the student to delete, on the grounds

that they were impossible in the contest for reasons which might be semantic, structural or a combination thereof.

2. Social scientific passage with missing words or parts of words to be inserted or completed.
3. Multiple-choice grammar items.
4. Literary passage with deliberate grammatical mistakes to be underlined and corrected by the student.

Test D

Tested reading comprehension. Consisted entirely of passages in different styles or registers of Spanish: 5 passages, each followed by a selection of the following types of question.

a) Choosing from several alternatives the most suitable synonym or translation for isolated words or structures, the choice being dependent on context.

b) Supplying a suitable synonym or translation for isolated words or structures.

c) Stating the acceptability of given translations of excerpts from the passages.

d) Judging the validity of given statements about the content of the passages.

Test E

Tested written skills.

Included:-

1. A letter in Spanish, to be reported as if by a biographer for the writer of the letter.
2. Précis, in Spanish, of an English passage.

APPENDIX 2

The following is a list of co-operating centres:

Universities (20)

Birmingham
Bradford
Bristol
Cardiff
Trinity College, Dublin
Edinburgh
Durham
Essex
Exeter
Liverpool
King's College, London
Queen Mary College, London
University College, London
Westfield College, London
Manchester
Newcastle
Nottingham
Salford
Sheffield
Southampton

Colleges (4)

Cambridgeshire College of Arts and Technology
Leeds College of Commerce
Trinity and All Saints' Colleges of Education
Wolverhampton Polytechnic

APPENDIX 3

Analysis of variance

Abbreviations:

SS = Sum of squares
DF = Degrees of freedom
MS = Mean square
VR = Variance ratio
SL = Significance level
NS = Not significant

GNA = A-level unknown
PLY = University language year only
PLS = Previous language study
PLSB = Bilinguals
PLSA = Previous study abroad
NPLSA = No previous study abroad
PLSNA = Details of PLS not known
PLSGE4 = More than 4 years study
PLSGL3 = Less than 3 years study

1. Test A

a) G.C.E.A-level

		SS	DF	MS	VR	SL
GNA v. Rest	7.8%	234.19	1	234.19	3.378	NS
PLY v. Grades		1.13	1	1.13	-	NS
Between grades	856.54	621.22	4	155.31	2.240	NS
Error		10260.00	148	69.32		
Total		11016.54	154	71.54		

		SS	DF	MS	VR	SL
b) PLS (abroad)						
PLSB v. Rest	36.6%	2470.20	1	2470.20	53.09	→ 0
PLSA v. NPLSA		1432.23	1	1432.23	30.59	→ 0
PLY v. PLSA / NPLSA v. PLSNA	4037.77	135.34	2	67.67	1.454	NS
Error		6978.77	150	46.53		
Total		11016.54	154	71.54		
c) PLS (duration)						
PLSB v. Rest	23.8%	2470.20	1	2470.20	44.11	→ 0
PLSGE4 v. PLSLE3		11.43	1	11.43	-	NS
PLY v. PLSGE4 / PLSLE3 v. PLSNA	2616.97	135.34	2	67.67	1.208	NS
Error		8399.57	150	56.00		
Total		11016.54	154	71.54		
2. Test B						
a) G.C.E. A-level						
PLY v. Rest	6.8%	979.41	1	979.41	2.423	NS
GNA v. Grades		193.65	1	193.65	-	NS
Between grades	6744.95	5571.89	4	1390.47	3.446	1%
Error		92557.82	229	404.18		
Total		99302.77	235			

b) PLS (abroad)		SS	DF	VR/MS	SL
PLSB v. Rest	16.3%	4745.25	1	13.18	0.1%
PLY v. PLSA / NPLSA / PLSNA	16153.84	1218.69	1	3.386	NS
PLSA v. NPLSA		9924.75	1	27.57	→ 0
PLSNA v. PLSA / NPLSA		265.15	1	-	NS
Error		83148.93	231	359.95	
Total		99302.77	235	422.56	

c) PLS (duration)		SS	DF	VR/MS	SL
PLSB v. Rest	6.4%	4745.25	1	11.78	0.1%
PLY v. PLSGE4 / PLSLE3 / PLSNA	6325.12	1218.69	1	3.028	NS
PLSGE4 v. PLSLE3		96.03	1	-	NS
PLSNA v. PLSGE4 / PLSLE3		265.15	1	-	NS
Error		92977.65	231	402.50	
Total		99302.77	235	422.56	

3. Test C

a) G.C.E. A-level		SS	DF	MS	VR/SL
Between categories	17.2%	12862.74	6	2143.79	8.801
Error		61873.90	254	243.60	0.1%
Total		74736.64	260	286.06	

b) PLS (abroad)				VR/MS	SL
PLSB v. Rest	15.5%	7636.38	1	30.97	→ 0
PLSA v. NPLSA		3965.19	1	16.08	0.1%
PLY v. PLSA / NPLSA v. PLSNA	11622.88	21.31	2	-	NS
Error		63113.76	256	246.54	
Total		74736.64	260	286.06	

c) PLS (duration)					
PLSB v. Rest	16.0%	7636.38	1	31.15	→ 0
PLSGE4 v. PLSLE3		4319.94	1	17.62	0.1%
PLY v. PLSGE4 / PLSLE3 v. PLSNA	11977.63	21.31	2	-	NS
Error		62759.01	256	245.15	
Total		74736.64	260	286.06	

4. Test D.

		SS	DF	MS	VR/SL
a) G.C.E. A-level					
Between categories	8.6%	2115.65	6	352.61	3.861
Error		22465.92	246	91.32	= 0.1%
Total		24581.57	252	97.55	
				VR/MS	SL
b) PLS (abroad)					
PLSB v. Rest	3.6%	188.17	1	1.970	NS
PLSA v. NPLSA		653.7	1	6.843	1%
PLY v. PLSA / NPLSA v. PLSNA	889.06	47.19	2	-	NS
Error		23692.51	248	95.53	
Total		24581.57	252	97.55	
c) PLS (duration)					
PLSB v. Rest		188.17	1	1.952	NS
PLSGE4 v. PLSLE3	2.9%	434.59	1	4.507	5%
PLY v. PLSGE4 / PLSLE3 v. PLSNA	669.95	47.19	2	-	NS
Error		23912.62	248	96.42	
Total		24581.57	252	97.55	

5. Test E

		SS	DF	MS	VR	SL
a) G.C.E. A-level						
Between grades	8.7%	1295.03	4	323.76	3.757	1%
Grades v. GNA/PLY	1346.44	51.54	2	25.71	-	NS
Error		14047.63	163	86.18		
Total		15394.09	169	91.09		
b) PLS (abroad)				VR/MS		
PLSB v. Rest	5.9%	601.24	1	6.846		1%
PLSA v. NPLSA	904.36	236.27	1	2.690		NS
PLY v. PLSA / NPLSA v. PLSNA		66.85	2	-		NS
Error		14490.73	165	87.82		
Total		15394.09	169	91.09		
c) PLS (duration)						
PLSB v. Rest	4.4%	601.24	1	6.738		2.5%
PLSGE4 v. PLSLE3	671.27	3.18	1	-		NS
PLY v. PLSGE4 / PLSLE3 v. PLSNA		66.85	2	-		NS
Error		14722.82	165	89.23		
Total		15394.09	169	91.09		